

Testing means from sampling populations with undefined labels

Florent AUTIN* (Aix-Marseille Université)
and Christophe POUET† (Ecole Centrale Marseille)

Abstract

We consider the problem of testing means from samples of two populations for which the labels are not defined with certainty. We show that this problem is connected to another one that is testing expected values of components of mixture-models from two data samples. The underlying mixture-model is associated with known varying mixing-weights. We provide a testing procedure that performs well. Then we point out the loss of performance of our method due to the mixing-effect by comparing its numerical performances to the Welch's t-test on means which would have been done if true labels were available.

Keywords: Asymptotic distribution, hypothesis testing, missing labels, mixture-models.

AMS Subject Classification: 62D05, 62F03, 62F05.

1 Introduction

In many cases, researchers can be interested in gathering information about two populations in order to compare them. In that setting, tests of significance are useful statistical tools for detecting a difference between two population parameters. Related application fields are numerous. Some examples are genetics, neuronal data analysis, medicine, biology, physics, chemistry, social sciences, among other fields.

Consider the Gaussian setting, for which each data of the two populations under study is assumed to follow a normal distribution. Let us recall that this assumption can be tested beforehand using a normality test, such as the well-known Shapiro-Wilk or Kolmogorov-Smirnov test, or it can be assessed graphically using a normal quantile plot. Comparisons between the means of the populations

*Address : C.M.I., 39 rue F. Joliot Curie, 13453 Marseille Cedex 13. Aix-Marseille Université. FRANCE. Email: autin@cmi.univ-mrs.fr

†Address : Ecole Centrale Marseille, 38 rue F. Joliot-Curie, 13451 Marseille Cedex 20. FRANCE. Email : cpouet@centrale-marseille.fr

are usually carried out by using t-statistic and lead to the well known Student's t-test or Welch's t-test (see Welch [9]).

These t-tests are popular because of their ease of use and their good performances. Moreover they are robust in the sense that they still perform well when the components are not really Gaussian, provided that the samples size are large enough. Nevertheless, these testing methods require to know the *label* of each data, that is the population each data is associated with. Unfortunately, researchers do sometimes not get this information. Indeed, one can imagine some cases where the labels of data are erroneous or uncertain, i.e. some data of each population do not deal with the population we want to compare. To give an example of such a situation with lack of information, we consider two populations - New York and California people - reduced to people that take bus/trolley bus or walk to go working. Focusing on working people that take bus/trolley bus, suppose you are interested to know whether travel time of people from New York is significantly different to the one from California from a sample of people where the place they live - New York or California - is available but the way of travel (the *label*) associated with each data in hand is not.

This kind of situation is the one we are interested in. Indeed, we want to address the problem of testing means of (sub)populations when the labels of data are uncertain. More precisely, we first propose to show that this testing problem can be reformulated as testing the expected values of components from two samples of independent mixture variables. In our study of real data, we shall assume that the mixing-weights are known. It means that proportions of people walking or using bus/trolley bus for each population (New York and California) are known, with respect to an auxiliary variable (age for instance). Then, we provide a testing procedure that takes into account this information on populations - and we discuss about its performances.

The testing procedure we propose is directly inspired from ideas in Autin and Pouet [1]. In this previous work, a nonparametric procedure has been proposed to test whether the densities of two independent samples of independent random variables result from the same mixture of components or not. The value of the test statistic requires to invert in some sense the mixing-weights operators of samples (see Definition 1) as a preliminary step to be calculated. This testing procedure was proved to be powerful since it is minimax over Besov spaces (more details are given in paragraph 3.1 in Autin and Pouet [1]). More focusing on practical purposes, we show that providing a testing procedure that incorporates combinatory ideas - provided that the mixing-weights are known - is quite relevant compared to a procedure usually used in classification.

Paper is organized as follows. In Section 2 we present the mixture-model we are interested in. Connection between the testing problem for which the labels of data are not certain and the problem of testing the expected value of the components involved in the mixture-model is provided. In Section 3, we present three testing procedures. The first one is the *Oracle Procedure* that uses Welch's t-test on data associated with the label of the components we want to test.

Of course this procedure is not tractable for the testing problem with lack of information about labels but it will be used as a benchmark to assess the loss of performances of the other tractable testing procedures. The second testing procedure we present is the *Expert Procedure* that uses Welch's t-test on data that are supposed to have, with probability larger than or equal to one half, the label associated with the components we focus on. The third and last procedure, namely the *Mixing Procedure*, uses combinatory properties leading to a new performing test. Section 4 deals with numerical experiments to point out the good performances of the Mixing Procedure - the one we suggest - compared to the Expert Procedure and to assess the loss of performances due to the mixing-effect compared to the Oracle Procedure. An application to real data is also presented whereas a brief conclusion is postponed in Section 5. Finally, the technical lemmas and the proposition we used to prove our main theoretical result (see Theorem 1) together with their proofs can be found in the appendix.

2 Model description and hypothesis testing problem

2.1 Mixture-models with varying mixing-weights

Let X_1, \dots, X_n be independent random variables such that, for any $1 \leq i \leq n$, the density of X_i on \mathbb{R} , denoted by f_{X_i} , is a mixture density with components p_1 and p_2 and mixing-weights $\omega_1(i)$ and $\omega_2(i)$, i.e.

$$f_{X_i} = \omega_1(i)p_1 + \omega_2(i)p_2.$$

We also introduce labels attached to X_1, \dots, X_n , denoted by u_1, \dots, u_n . This point of view is one interpretation of mixture-models among others (see Section 1.4 in McLachlan and Peel [4]). The main difference lies in considering varying mixing-weights in our model. This point is very important (see Autin and Pouet[1]). Therefore our model cannot be described as a mixture-model in the usual sense.

Similarly to the sample X_1, \dots, X_n , we consider a sample of independent random variables Y_1, \dots, Y_n such that, for any $1 \leq i \leq n$ the density of Y_i on \mathbb{R} , denoted by f_{Y_i} , is a mixture density with components p'_1 and p'_2 and mixing-weights $\omega'_1(i)$ and $\omega'_2(i)$, i.e.

$$f_{Y_i} = \omega'_1(i)p'_1 + \omega'_2(i)p'_2.$$

We also introduce labels attached to Y_1, \dots, Y_n , denoted by v_1, \dots, v_n and we assume that this second sample is independent from the first one.

If t denotes the transpose operator, the two mixture-models we have just introduced can be rewritten in a simpler way as follows:

$$\mathbf{f}_X = \Omega_X \mathbf{p} \quad \text{and} \quad \mathbf{f}_Y = \Omega_Y \mathbf{p}', \quad (1)$$

Table 1: Populations weights (and sizes) with respect to age

	Bus/trolleybus	Walk
New York over 21 y.o.	51.93% (4313)	48.07% (3993)
New York under 20 y.o.	34.65% (306)	65.35% (577)
California over 21 y.o.	57.4% (4479)	42.6% (3324)
California under 20 y.o.	42.77% (497)	57.23% (665)

with,

- $\mathbf{f}_X = {}^t(f_{X_1}, \dots, f_{X_n})$, $\mathbf{f}_Y = {}^t(f_{Y_1}, \dots, f_{Y_n})$,
- $\mathbf{p} = {}^t(p_1, p_2)$, $\mathbf{p}' = {}^t(p'_1, p'_2)$,
- $\Omega_X = (\omega_l(i))_{i,l}$, $\Omega_Y = (\omega'_l(i))_{i,l}$.

Definition 1 *The matrices Ω_X and Ω_Y involved in the model (1) are called the mixing-weights operators.*

Definition 2 *Any mixture-model (1) such that Ω_X and Ω_Y are full rank matrices is called mixture-model with varying mixing-weights.*

2.2 Example of modeling with mixture-models

Let us illustrate this theoretical set-up with the example cited in the introduction. The random variables X_1, \dots, X_n correspond to the travel times of people in the state of New York and the random variables Y_1, \dots, Y_n to travel times in the state of California. The labels are the ways of transportation to go working and can be either *Bus/trolley bus* (label 1) or *Walk* (label 2). The last step to complete the mixture model is to describe the mixing-weights for each observation. In each state the mixing-weights strongly depend on the age (over 21 or under 20 years old (y.o.)). Table 1 illustrates this fact.

This table leads to the following mixing-weights:

$$\begin{aligned}
 \omega_1(i) &= 0.5193, \omega_2(i) = 0.4807 \\
 &\quad \text{if the person } i \text{ is over 21 y.o. and lives in New York,} \\
 \omega_1(i) &= 0.3465, \omega_2(i) = 0.6535 \\
 &\quad \text{if the person } i \text{ is under 20 y.o. and lives in New York,} \\
 \omega'_1(i) &= 0.574, \omega'_2(i) = 0.426 \\
 &\quad \text{if the person } i \text{ is over 21 y.o. and lives in California,} \\
 \omega'_1(i) &= 0.4277, \omega'_2(i) = 0.5723 \\
 &\quad \text{if the person } i \text{ is over 21 y.o. and lives in California.}
 \end{aligned}$$

The reader can legitimately wonder why the age is assumed to be known and not the ways of transportation to go working. One can think about at least two good reasons. The first one is an a priori reason. The survey would be rather lengthy if all interesting variables were included. Therefore the survey is restricted to a small set of informative variables strongly linked with the interesting variables. Moreover these informative variables can be chosen as objective as possible and easily recordable. This can be called planned missing values (see Graham [3]). The other reason is an a posteriori one. During the data analysis of a survey, researchers are often confronted with new hypotheses to test. In many situations, the relevant variables have not been recorded and researchers have to plan a new survey which includes these new variables in order to check these hypotheses. This leads to a waste of time and money. Our testing problem on means from data with undefined labels can be associated with the testing problem (2) in the mixture-model (1). Indeed, it corresponds to a testing problem on means for which labels of data are unavailable: the only information on the X_i 's label (resp. Y_i 's label) is the probability $\omega_l(i)$ (resp. $\omega'_l(i)$) that it corresponds to l , for any $l \in \{1, 2\}$. In other words, the added information on subpopulations we get is the knowledge of the mixing-weights operators.

2.3 Hypothesis testing problem

We recall that two data samples $\mathbf{X} = {}^t(X_1, \dots, X_n)$ and $\mathbf{Y} = {}^t(Y_1, \dots, Y_n)$ are considered. For a chosen label $l \in \{1, 2\}$, we are interested in testing whether components p_l and p'_l have the same expected value or not. We want to address this problem in a general context that is: the parameters of variance σ_k^2 and $\sigma_k'^2$ of the components p_k and p'_k are unknown whatever $k \in \{1, 2\}$.

For a fixed $l \in \{1, 2\}$, when respectively denoting by m_l and m'_l the expected value of the components we focus on, the testing problem we consider is lying on the two following hypotheses:

$$\text{the null hypothesis} \quad \mathcal{H}_0 : m_l = m'_l, \tag{2}$$

$$\text{the alternative hypothesis} \quad \mathcal{H}_1 : m_l \neq m'_l.$$

We recall that providing a procedure to solve the testing problem (2) means giving a decision rule (or test) $\Delta \in \{0, 1\}$ that relies on the value of a measurable function T (test statistic) of X_1, \dots, X_n and Y_1, \dots, Y_n .

As usual, $\Delta = 1$ will mean deciding \mathcal{H}_1 whereas $\Delta = 0$ will mean deciding \mathcal{H}_0 .

3 Description of testing procedures

In this section we introduce the testing procedures we are interested in.

3.1 Oracle test: Δ_o

The first testing procedure we present is called the *Oracle Procedure*. This is a two steps procedure. First step consists in recovering the true labels of data. Second step lies on using the Welch's t-test on data with label l in order to know whether m_l and m'_l can be judged as different. This test cannot be used in our context where the true labels are unknown but it will be used as a benchmark when comparing the performances of the other testing procedures. It corresponds to the procedure proposed by the *oracle*: any statistician having information on labels.

Here we describe into details the Oracle Procedure. Let us denote by

- $n_l = \sum_{i=1}^n \mathbf{1}\{u_i = l\}$ and $n'_l = \sum_{i=1}^n \mathbf{1}\{v_i = l\}$,
- $\bar{X}^{(l)} = \frac{1}{n_l} \sum_{i=1}^n X_i \mathbf{1}\{u_i = l\}$ and $\bar{Y}^{(l)} = \frac{1}{n'_l} \sum_{i=1}^n Y_i \mathbf{1}\{v_i = l\}$,
- $\hat{\sigma}_l^2 = \frac{1}{n_l} \sum_{i=1}^n (X_i - \bar{X}^{(l)})^2 \mathbf{1}\{u_i = l\}$ and $\hat{\sigma}'^2_l = \frac{1}{n'_l} \sum_{i=1}^n (Y_i - \bar{Y}^{(l)})^2 \mathbf{1}\{v_i = l\}$.

The Oracle test Δ_o lies on the test statistic T_o defined as follows

$$T_o := \frac{|\bar{X}^{(l)} - \bar{Y}^{(l)}|}{\sqrt{\frac{\hat{\sigma}_l^2}{n_l} + \frac{\hat{\sigma}'^2_l}{n'_l}}}.$$

Under the null hypothesis, the asymptotic law of T_o is known to be the Standard Gaussian one, namely $\mathcal{N}(0, 1)$. Hence, $\Delta_o = \mathbf{1}\{T_o > q_r\}$ is a test with asymptotically type I error equal to r ($0 < r < 1$), where q_r is the quantile of order $1 - \frac{r}{2}$ of the Standard Gaussian law.

3.2 Expert test: Δ_e

The testing procedure we describe now is lying on a method used in classification. It is a two steps procedure. The first step consists in allocating label l to any data X_i such that $\omega_l(i) \geq \frac{1}{2}$ and to any data Y_j such that $\omega'_l(j) \geq \frac{1}{2}$. The second step consists in using the Welch's t-test on the two subsamples of data that have been assigned to label l to know whether m_l and m'_l can be judged as different. Notice that it means that the Welch's t-test is done on data having possible wrong labels.

Put

- $n_{l,e} = \sum_{i=1}^n \mathbf{1}\{\omega_l(i) \geq \frac{1}{2}\}$ and $n'_{l,e} = \sum_{i=1}^n \mathbf{1}\{\omega'_l(i) \geq \frac{1}{2}\}$,

$$\begin{aligned}
- \bar{X}_e^{(l)} &= \frac{1}{n_{l,e}} \sum_{i=1}^n X_i \cdot \mathbf{1} \left\{ \omega_l(i) \geq \frac{1}{2} \right\} \text{ and } \bar{Y}_e^{(l)} = \frac{1}{n'_{l,e}} \sum_{i=1}^n Y_i \cdot \mathbf{1} \left\{ \omega'_l(i) \geq \frac{1}{2} \right\}, \\
- \hat{\sigma}_{l,e}^2 &= \frac{1}{n_{l,e}} \sum_{i=1}^n (X_i - \bar{X}_e^{(l)})^2 \mathbf{1} \left\{ \omega_l(i) \geq \frac{1}{2} \right\}, \\
- \hat{\sigma}_{l,e}'^2 &= \frac{1}{n'_{l,e}} \sum_{i=1}^n (Y_i - \bar{Y}_e^{(l)})^2 \mathbf{1} \left\{ \omega'_l(i) \geq \frac{1}{2} \right\}.
\end{aligned}$$

The Expert test Δ_e relies on the test statistic T_e defined as follows

$$T_e := \frac{|\bar{X}_e^{(l)} - \bar{Y}_e^{(l)}|}{\sqrt{\frac{\hat{\sigma}_{l,e}^2}{n_{l,e}} + \frac{\hat{\sigma}_{l,e}'^2}{n'_{l,e}}}}.$$

Then, the decision rule is done by putting $\Delta_e = \mathbf{1}\{T_e > q_r\}$.

3.3 Mixing test: Δ_m

The last testing procedure we propose is inspired from some ideas provided in Autin and Pouet [1]. Using combinatory methods, it proposes to invert in some sense the mixing-weights operators so as to provide a new test that will be proved to perform well. Let us describe this new testing procedure into details.

Let us denote by A_X and A_Y the matrices with n lines and 2 columns satisfying

$${}^t\Omega_X A_X = {}^t\Omega_Y A_Y = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}. \quad (3)$$

Notations: For any $(i, l) \in \{1, \dots, n\} \times \{1, 2\}$, we denote respectively by $a_l(i)$ (resp. $a'_l(i)$) the entries of A_X (resp. A_Y) associated with line i and column l .

Following Maiboroda [5] or Pokhyl'ko [7], solutions of equations (3) are given by

$$\begin{aligned}
a_l(i) &= \frac{n}{\det({}^t\Omega_X \Omega_X)} \sum_{k=1}^2 (-1)^{l+k} \gamma_{lk}^{(X)} \omega_k(i), \\
a'_l(i) &= \frac{n}{\det({}^t\Omega_Y \Omega_Y)} \sum_{k=1}^2 (-1)^{l+k} \gamma_{lk}^{(Y)} \omega'_k(i),
\end{aligned}$$

where $\gamma_{lk}^{(X)}$ and $\gamma_{lk}^{(Y)}$ are respectively the minor (l, k) of the matrix ${}^t\Omega_X \Omega_X$ and of the matrix ${}^t\Omega_Y \Omega_Y$.

For any $l \in \{1, 2\}$, (m_l, m'_l) can be estimated by the method of moments when using estimators (\hat{m}_l, \hat{m}'_l) defined as follows:

$$\begin{aligned} (\hat{m}_l, \hat{m}'_l) &= \left(\langle A_x^{(l)}, X \rangle_n, \langle A_y^{(l)}, Y \rangle_n \right) \\ &:= \left(\frac{1}{n} \sum_{i=1}^n a_l(i) X_i, \frac{1}{n} \sum_{i=1}^n a'_l(i) Y_i \right), \end{aligned}$$

provided that $A_x^{(l)}$ and $A_y^{(l)}$ respectively denote the l -th column-vector of matrices A_x and A_y .

The Mixing test Δ_m lies on the test statistic T_m defined by:

$$T_m := \frac{|\hat{m}_l - \hat{m}'_l|}{\sqrt{\hat{\mathbb{V}}_n^{(l)}}}, \quad (4)$$

where $\hat{\mathbb{V}}_n^{(l)}$ is the estimated variance of $\hat{m}_l - \hat{m}'_l$, that is

$$\hat{\mathbb{V}}_n^{(l)} = \frac{1}{n^2} \sum_{i=1}^n \left[a_l^2(i) (X_i - \omega_1(i)\hat{m}_l - \omega_2(i)\hat{m}_l)^2 + a_l'^2(i) (Y_i - \omega'_1(i)\hat{m}'_l - \omega'_2(i)\hat{m}'_l)^2 \right].$$

Remark 1 *As discussed in Autin and Pouet [1], for any $l \in \{1, 2\}$ the random variable \hat{m}_l (resp. \hat{m}'_l) is a good estimator for m_l (resp. m'_l). Hence if the distance between \hat{m}_l and \hat{m}'_l is judged too large, the rejection of the null hypothesis \mathcal{H}_0 looks better. This idea motivates the choice of the test statistic T_m we defined above.*

Under the null hypothesis \mathcal{H}_0 , the asymptotic law of T_m is known, according to the following Theorem.

Theorem 1 *Let $l \in \{1, 2\}$. Assume that*

- *the components within the mixture-model (1) have moments with order 4,*
- *the mixing-weights of the mixture-model (1) are such that*

$$\lim_{n \rightarrow +\infty} \frac{\sup_{i=1, \dots, n} a_l^2(i)}{\sum_{i=1}^n a_l^2(i)} = \lim_{n \rightarrow +\infty} \frac{\sup_{i=1, \dots, n} a_l'^2(i)}{\sum_{i=1}^n a_l'^2(i)} = 0. \quad (5)$$

Then, under the null hypothesis \mathcal{H}_0 , the law of T_m is asymptotically the Standard Gaussian one, i.e.

$$T_m \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad (6)$$

Hence, $\Delta_m = \mathbf{1}\{T_m > q_r\}$ is a test with asymptotically type I error equal to r ($0 < r < 1$).

Remark 2 *A wide range of mixing-weights of the mixture-model (1) satisfy the condition (5). Examples of such mixing-weights are given in (12) of Section 4.*

Proof:

To prove Theorem 1, notice that it suffices to prove that for any $l \in \{1, 2\}$

$$\begin{aligned} 1. & \frac{\frac{1}{n} \sum_{i=1}^n a_l(i) X_i - m_l}{\sqrt{\frac{1}{n^2} \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i) \hat{m}_1 - \omega_2(i) \hat{m}_2)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \\ 2. & \frac{\frac{1}{n} \sum_{j=1}^n a'_l(i) Y_j - m'_l}{\sqrt{\frac{1}{n^2} \sum_{j=1}^n a_l'^2(i) (Y_j - \omega'_1(j) \hat{m}'_1 - \omega'_2(i) \hat{m}'_2)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \end{aligned}$$

Because of the independence between the two samples and the fact that under the null hypothesis \mathcal{H}_0 , $m_l = m'_l$. Since these two results of convergence can be proved by an analogous way, we only focus on proving the first one that can be rewritten as follows for any $l \in \{1, 2\}$:

$$\frac{\sum_{i=1}^n a_l(i) (X_i - \omega_1(i) m_1 - \omega_2(i) m_2)}{\sqrt{\sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i) \hat{m}_1 - \omega_2(i) \hat{m}_2)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Denote, for any $n \in \mathbb{N}^*$, any $l \in \{1, 2\}$ and any $1 \leq i \leq n$

$$B_n^{(l)} = \sum_{i=1}^n a_l^2(i) \mathbb{E} \left[(X_i - \omega_1(i) m_1 - \omega_2(i) m_2)^2 \right], \quad (7)$$

$$\hat{B}_n^{(l)} = \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i) \hat{m}_1 - \omega_2(i) \hat{m}_2)^2. \quad (8)$$

From Proposition 1,

$$\frac{\sum_{i=1}^n a_l(i) (X_i - \omega_1(i) m_1 - \omega_2(i) m_2)}{\sqrt{B_n^{(l)}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

In the sequel, we aim at proving that a same kind of result holds when replacing parameter $B_n^{(l)}$ by the estimator $\hat{B}_n^{(l)}$. In other words, the result we want to prove is the following:

$$\frac{\sum_{i=1}^n a_l(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2)}{\sqrt{\hat{B}_n^{(l)}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad (9)$$

From Slutsky theorem, it suffices to prove that estimator $\hat{B}_n^{(l)}$ of $B_n^{(l)}$ is consistent. We propose to divide the proof of this consistency into two steps. First we prove

$$\frac{\sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^2}{B_n^{(l)}} \xrightarrow{Prob} 1. \quad (10)$$

The second step consists in replacing m_1 and m_2 by their consistent estimators \hat{m}_1 and \hat{m}_2 and in checking that the convergence in probability still holds.

From this point we need more assumptions, that is to say the existence of the fourth order moment for p_1 and p_2 .

Let us prove the first step. We apply Bienayme-Chebyshev inequality, for any $\epsilon > 0$:

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^2 - B_n^{(l)} \right| > B_n^{(l)} \epsilon \right) \\ & \leq \frac{\sum_{i=1}^n a_l^4(i) \mathbb{V}ar \left((X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^2 \right)}{(B_n^{(l)} \epsilon)^2} \\ & \leq \frac{\sum_{i=1}^n a_l^4(i) \mathbb{E} \left[(X_i - \mathbb{E}(X_i))^4 \right]}{(B_n^{(l)} \epsilon)^2} \\ & \leq \frac{\sup_{j=1, \dots, n} a_l^2(j)}{B_n^{(l)}} \frac{\sum_{i=1}^n a_l^2(i) C(m_1, m_2, p_1, p_2)}{B_n^{(l)} \epsilon^2} \\ & \leq \frac{\sup_{j=1, \dots, n} a_l^2(j)}{B_n^{(l)}} \frac{(\min(\sigma_1^2, \sigma_2^2))^{-1} C(m_1, m_2, p_1, p_2)}{\epsilon^2}. \end{aligned}$$

Last inequalities are obtained by using Lemma 4 and Lemma 2. The right part of the last inequality is the product of two terms. The left one tends to 0 when

n goes to infinity because of assumption (5). The right one is a constant that only depends on ϵ and the parameters of p_1 and p_2 . When considering the limit in infinity with respect to n , we conclude that property (10) holds.

We end by proving the second step. We have

$$\begin{aligned}
& \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)\hat{m}_1 - \omega_2(i)\hat{m}_2)^2 \\
&= \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^2 \\
&+ 2 \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2) (\omega_1(i)(m_1 - \hat{m}_1) + \omega_2(i)(m_2 - \hat{m}_2)) \\
&+ \sum_{i=1}^n a_l^2(i) (\omega_1(i)(m_1 - \hat{m}_1) + \omega_2(i)(m_2 - \hat{m}_2))^2.
\end{aligned}$$

The first term is exactly the one appearing in the first step and also converges to 1 in probability when divided by $B_n^{(l)}$. We turn to the second term. Cauchy-Schwarz inequality entails that

$$\begin{aligned}
& \left| \sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2) (\omega_1(i)(m_1 - \hat{m}_1) + \omega_2(i)(m_2 - \hat{m}_2)) \right| \\
&\leq \sqrt{\sum_{i=1}^n a_l^2(i) (X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^2} \\
&\quad \times \sqrt{2(m_1 - \hat{m}_1)^2 \sum_{i=1}^n a_l^2(i) \omega_1^2(i) + 2(m_2 - \hat{m}_2)^2 \sum_{i=1}^n a_l^2(i) \omega_2^2(i)}.
\end{aligned}$$

When divided by $B_n^{(l)}$, the first term of the righthand-side converges to 1 in probability: it is the result of the first step. By using Lemma 2, one gets

$$\max \left(\sum_{i=1}^n a_l^2(i) \omega_1^2(i), \sum_{i=1}^n a_l^2(i) \omega_2^2(i) \right) \leq B_n^{(l)} (\min(\sigma_1^2, \sigma_2^2))^{-1}.$$

Hence the second term of the right-hand side of the inequality converges to 0 in probability when divided by $B_n^{(l)}$ because of the consistency of estimators \hat{m}_l (see Lemma 5). Hence second term we are interested in converges to 0 in probability when divided by $B_n^{(l)}$. We can proceed in the same way in order to prove that the third term converges to 0 in probability when divided by $B_n^{(l)}$.

So, we have just proved that

$$\frac{\hat{B}_n^{(l)}}{B_n^{(l)}} \xrightarrow{Proba} 1. \tag{11}$$

We conclude that the exact variance $B_n^{(l)}$ can be replaced by the consistent estimator $\hat{B}_n^{(l)}$ for the result of convergence. In other words, the property (9) holds.

4 Numerical experiments

4.1 Numerical performances of the Mixing-test

In this section we provide numerical experiments and we discuss about the performances of our testing procedure. What we often expect is a gain of performance of the test Δ_m - that is to say a smaller type II error when the type I error is chosen to be $r = 0.05$ - comparatively to the test Δ_e . Without loss of generality, we suppose that n is even.

We consider the Gaussian setting and we assume in this section that the mixing-weights operators Ω_x and Ω_y have the following form:

$$\Omega_x = \begin{pmatrix} \alpha & 1-\alpha \\ \dots & \dots \\ \alpha & 1-\alpha \\ 1-\beta & \beta \\ \dots & \dots \\ 1-\beta & \beta \end{pmatrix} \text{ and } \Omega_y = \begin{pmatrix} \alpha' & 1-\alpha' \\ \dots & \dots \\ \alpha' & 1-\alpha' \\ 1-\beta' & \beta' \\ \dots & \dots \\ 1-\beta' & \beta' \end{pmatrix}, \quad (12)$$

where $\frac{n}{2}$ data from \mathbf{X} (resp. \mathbf{Y}) deal with the couple of mixing-weights $(\alpha, 1-\alpha)$ (resp. $(\alpha', 1-\alpha')$) and the other $\frac{n}{2}$ data from \mathbf{X} (resp. \mathbf{Y}) deal with the couple of mixing-weights $(1-\beta, \beta)$ (resp. $(1-\beta', \beta')$). Suppose now that our testing problem is dealing with the first component, i.e. $l = 1$ and that Ω_x and Ω_y are full rank matrices, i.e. $\alpha + \beta \neq 1$ and $\alpha' + \beta' \neq 1$.

4.1.1 Mixing-test versus Expert-test

In this paragraph we provide a motivation for the use of our testing procedure Δ_m . For the sake of simplicity we suppose that $\alpha = \beta$ and that $\alpha' = \beta'$. For any value of $(\alpha, \alpha') \in]\frac{1}{2}, 1]^2$, there are many situations where the performance of the Expert test is quite bad even if the numbers of observations n is large.

- Dealing with two components with equal expected value, Δ_e can most of time detect a difference between these components (wrong decision) whereas our test doesn't. For instance, suppose that $m_1 = m'_1$ and that m'_2 is large away from m_2 as $\alpha = \alpha'$. Since $\mathbb{E}(\bar{X}_e^{(1)}) \neq \mathbb{E}(\bar{Y}_e^{(1)})$, using Δ_e to detect equality between components m_1 and m'_1 would be a very bad choice in that context. For n large enough, it would imply that $T_e > t_r$ with high probability. Hence, the wrong decision \mathcal{H}_1 may often be done.

An example of such a situation is given here in the case where $\alpha = \alpha' = 0.9$. Consider the testing problem (2) and suppose $\sigma_1 = \sigma'_1 = \sigma_2 = \sigma'_2 = 1$ and

Table 2: Percentage of wrong decisions by Δ_e

δ / n	100	200	500	1000	2000
0.5	0.057	0.064	0.086	0.121	0.191
1	0.074	0.098	0.172	0.302	0.521
2	0.126	0.210	0.462	0.749	0.963
3	0.188	0.350	0.722	0.950	0.999

Table 3: Percentage of correct decisions by Δ_m

n	500	1000	2000	3000	4000	5000	6000
Δ_m	0.146	0.242	0.438	0.595	0.718	0.810	0.879

that $m_1 = m'_1 = 0$, $m_2 = 1$ and $m'_2 = m_2 + \delta$. For varying values of n , δ and 40 000 repetitions of Δ_e with $r = 0.05$, we give the percentage of wrong decisions \mathcal{H}_1 in Table 2.

Notice that the percentage of wrong decisions by Δ_e turns up as n grows up and can be quite important if m'_2 is sufficiently far away from m_2 . Most of time, the expert detects a difference between the components m_1 and m'_1 but there is not in that context. Comparatively speaking, the percentage of wrong decisions by Δ_m is around 0.05.

- Most of time Δ_e fails to detect a difference between two components with different expected value whereas our test doesn't. For instance, suppose that $m_1 \neq m'_1$ and that

$$m_2 \approx (1 - \alpha)^{-1} (\alpha' m'_1 + (1 - \alpha') m'_2 - \alpha m_1).$$

Since

$$\mathbb{E}(X_i) \approx \mathbb{E}(Y_i), \quad \text{for any } 1 \leq i \leq \frac{n}{2},$$

using Δ_e to detect the difference between m_1 and m'_1 would be a very bad choice in that context. Indeed, according to the law of large numbers, with high probability - that increases as n goes up - $\bar{X}_e^{(1)}$ and $\bar{Y}_e^{(1)}$ would be very close to each other. It would imply that $T_e \leq 1.96$ with high probability. True decision \mathcal{H}_1 would be taken only in 5 % of cases.

An example of such a situation is given here in the case $\alpha = \alpha' = 0.9$. Consider the testing problem (2) and suppose that $\sigma_1 = \sigma'_1 = \sigma_2 = \sigma'_2 = 1$ and that $m_1 = 0$, $m'_1 = 0.1$, $m_2 = 1$ and $m'_2 = 2$. For varying values of n and 40 000 repetitions of Δ_m with $r = 0.05$, we give the percentage of correct decisions from Δ_m in Table 3.

As expected, the percentage of correct decisions by Δ_m goes up as n grows up. But it is not the case for the percentage of correct decisions by Δ_e

Table 4: Empirical Power of Δ_o and Δ_m

Test / n	500	1000	2000	3000	4000	5000	6000
Δ_o	0.200	0.349	0.609	0.783	0.886	0.942	0.973
Δ_m	0.149	0.245	0.427	0.585	0.704	0.798	0.868

Table 5: Empirical Power \mathcal{P}_m of Δ_m for varying $\alpha = \alpha'$

δ / α	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
0.1	0.071	0.092	0.125	0.161	0.196	0.239	0.275	0.317	0.353
0.2	0.130	0.226	0.353	0.484	0.603	0.700	0.780	0.837	0.883
0.3	0.238	0.446	0.659	0.816	0.912	0.960	0.983	0.999	1
0.4	0.374	0.671	0.882	0.938	0.993	1	1	1	1

which are always around 0.05. Most of time, the expert is unable to detect the difference between the components in that context.

Finally we conclude that it is better to choose Δ_m for the problem we are interested in.

4.1.2 Mixing-test versus Oracle-test

In this paragraph we compare the *empirical powers* of Δ_m to the Oracle test Δ_o ones, when taking $r = 0.05$ and the same parameters as the last example. We recall that the empirical power of any test Δ corresponds to the numerical evaluation of the probability to correctly decide \mathcal{H}_1 , according to Δ .

According to Table 4 we remark that the bigger n the better the powers of Δ_o and of Δ_m . Moreover we note that the empirical power of the Mixing test is not bad when comparing to the Oracle test.

In Table 5 we give the empirical power \mathcal{P}_m of Δ_m measured in samplings of size $n = 1000$ in the case where $m_1 = 0$, $m'_1 = \bar{\delta}$, $m_2 = 1$ and $m'_2 = 0$.

As expected, looking at Table 5,

- quantity \mathcal{P}_m depends on the intrinsic difficulty of the problem. Indeed the larger the absolute value of quantity $\bar{\delta} := m'_1 - m_1$, the easier the problem of detection and so the more powerful the test,
- the larger the degree of certainty α the better the power of Δ_m . This is due to the fact that the expectation of the number of wrong labels considered by the Expert Procedure grows up as α goes down.

Table 6: Percentage of correct decisions by Δ_m

$(\alpha, \alpha'), / n$	100	200	500
(0.90, 0.60)	0.153	0.254	0.533
(0.80, 0.70)	0.311	0.536	0.896
(0.75, 0.75)	0.332	0.564	0.918

4.1.3 Comparisons on performances of Δ_m for varying values of (α, α')

As previously discussed, we expect that the better the degree of certainties of the expert, the better the performance of the test Δ_m . This statement is highlighted here when considering the same parameters as before, $\bar{\delta} = 0.5$ and many choices of couple (α, α') . For each choice of (α, α') done, we provide in Table 6 the empirical power \mathcal{P}_m of our test Δ_m that is the percentage of correct detection of a difference between m_1 and m'_1 .

Interpretation of the results presented in Table 6 goes in the same way as Autin and Pouet [1]: the bigger the smallest eigenvalue of both operators ${}^t\Omega_x\Omega_x$ and ${}^t\Omega_y\Omega_y$ that is $\lambda_{min} = \frac{1}{2}(1 - 2\min(\alpha, \alpha')(1 - \min(\alpha, \alpha')))$ the better the power of our test Δ_m . Note that, the larger the minimum value between $\alpha \in]\frac{1}{2}, 1[$ and $\alpha' \in]\frac{1}{2}, 1[$ the bigger λ_{min} .

4.1.4 Brief conclusion

Let us summarize the main facts. First, in some cases experts can be completely wrong because of the overall design, that is to say the link between the means of the components and the mixing-weights. This is a serious issue for the Expert test. The results become worse and worse as the sample size increases. The test adapted to the varying mixing-weights that we propose does not suffer from this drawback. The second fact is the good behavior of our test compared to the Oracle test. Although it is behind, the power is quite acceptable. The last important fact which has already been stressed by Autin and Pouet [1] is the effect of the mixing-weights. It is known a priori thanks to the smallest eigenvalue of the operators ${}^t\Omega_x\Omega_x$ and ${}^t\Omega_y\Omega_y$. This point is important as the statistician can act in order to counter to this effect, e.g. he can improve the accuracy of the expert system giving the mixing-weights or increase the sample sizes.

4.2 Application to real data

In this section we apply our methodology to real data and we discuss about the results.

Table 7: Description of the population

	NY	CA
Total	9189	8935
Over 21	90.39% (8306)	87.04% (7803)
Under 20	9.61% (883)	12.96% (1162)
Walk	49.73% (4570)	44.5% (3989)
Bus/trolley bus	50.27% (4619)	55.5% (4979)

Table 8: One-way analysis of travel time (in minutes)

	Walk	Bus/trolley bus	Walk and Bus/trolley bus
NY	12.25 (12.18)	47.26 (28.79)	29.85 (28.23)
CA	11.23 (12.23)	45.12 (28.84)	30.04 (28.49)

4.2.1 Description of the data

We have selected data from U.S. Census Bureau website, more precisely PUMS 2006 (see [8]). We are interested in comparing travel time of people living either in the state of New York (abbreviated in NY) or either in the state of California (abbreviated in CA). Two ways of transportation have been kept: Bus/trolley bus and Walk. We have also kept a variable linked to age as it will be useful for the mixture-model with varying mixing weights. This variable records the fact that a person is over 21 years old or under 20 years old.

Here are few facts to roughly describe the PUMS sample. Table 7 gives one-level information.

In Table 8 we compute the mean and the standard deviation (in parentheses) of the travel time according to the categorical variable means of transportation to go working.

As it can be seen in Table 8 there might be no difference between New York and California. Nevertheless if the means of transportation is unavailable, it will be perilous to decide when considering the whole sample without any other information. Indeed as shown in Table 7, the difference between New York and California is decreased because of the structure of the population (less people under 20 years old in New York).

4.2.2 Methodology

We assume in the sequel that the information about the way of transport (labels) are unavailable at the microdata level. We are going to apply the test Δ_m adapted to the varying mixing-weights mixture-model. The age variable is the only auxiliary information available at the microdata level that permits to get the mixing-weights to our mixture-model (1).

Table 9: Bus/trolleybus decisions

	Decision $n = 1000$	p-value
Oracle test	not rejected	0.24
Expert test	not rejected	0.42
Mixing test	not rejected	0.11

Table 10: Walk decisions

	Decision $n = 1000$	p-value
Oracle test	not rejected	0.23
Expert test	rejected	0.04
Mixing test	not rejected	0.48

For comparison purpose we have also applied the so-called Expert test. The type I error is chosen to be 0.1.

According to the notations we introduced in (12) and to Table 1,

$$(\alpha, \beta) = (0.5193, 0.6535) \quad (\alpha', \beta') = (0.574, 0.5723). \quad (13)$$

We consider the following sample: 500 persons over 21 and 500 persons under 20 were randomly sampled in each state ($n = 1000$).

We applied three testing procedures:

1. Oracle test,
2. Expert test,
3. Mixing test.

First we test the equality of the averages when the ways of transportation to work is Bus/trolley bus (label 1) in Table 9. In this case, the other means of transportation to work is considered as a nuisance parameter.

Next we reverse the set-up. We test the equality of the averages when the means of transportation to work is Walk (label 2) in Table 10. Bus/trolley bus is now a nuisance parameter.

4.2.3 A tough situation

Here we are also interested in comparing travel time of people living either in the state of New York or either in the state of Illinois (abbreviated in IL). Data come from U.S. Census Bureau [8]. Two ways of transportation to work have been kept: Bus/trolley bus or Railroad. We have also kept the gender variable as it will be useful for the varying mixing-weights mixture-model. As it will be seen, the situation is much more involved compared to the one in the previous section.

Table 11: Description of the population

	NY	IL
Total	6974	2899
Men	46.5% (3247)	48.2% (1398)
Women	53.5% (3727)	51.8% (1501)
Bus/trolley bus	66.2 % (4619)	58.4% (1692)
Railroad	33.8% (2355)	41.6% (1207)

Table 12: Mixing-weights

	Bus/trolley bus	Railroad
NY men	55.8 % (1813)	44.2% (1434)
NY women	75.3 % (2806)	24.7% (921)
IL men	50.8 % (710)	49.2% (688)
IL women	65.4 % (982)	34.6% (519)

Here are few facts to roughly describe the PUMS sample. Table 11 gives one-level information.

The mixing-weights depend on the gender as illustrated in Table 12.

According to the notations we introduced in (12) and Table 12,

$$(\alpha, \beta) = (0.558, 0.247) \quad (\alpha', \beta') = (0.508, 0.346). \quad (14)$$

In Table 13 we compute the mean and standard deviation (in parentheses) of the travel time according to the categorical variable way of transportation to work.

Once again the difference in travel time is decreased if we consider the entire population. This is due to its structure. As there are more men and women who use railroad in Illinois, the general average of travel time is increased. This is reverse in New York.

In Table 14 we test the equality of the averages when the ways of transportation to work is Bus/trolley bus.

In Table 15 we reverse the set-up and we test the equality of the averages when the way of transportation to work is Walk.

Table 13: One-way analysis of travel time

	Bus/trolley bus	Railroad	Bus/trolley bus and Railroad
New York	47.3 (28.8)	71 (30)	55.3 (31.3)
Illinois	41.8 (26.4)	63.1 (25.7)	50.7 (28.2)

Table 14: Bus/trolleybus decision

	Decision $n = 1000$	p-value
Oracle test	not rejected	0.19
Expert test	not rejected	0.13
Mixing test	not rejected	0.75

Table 15: Walk decision

	Decision $n = 1000$	p-value
Oracle test	rejected	0.05
Expert test	non-available	non-available
Mixing test	not rejected	0.12

5 Conclusion

From our point of view, one of the most interesting point is the usefulness of the varying mixing-weights model. It is a versatile model that can be used in many situations with missing microdata but aggregated information. The application treated exemplifies the modeling.

The second take-away message is the excellent performances of the Mixing test we propose. They can be guessed a priori thanks to the smallest eigenvalue of operators involved within the mixture-model. These nice performances were showed both theoretically and numerically.

To conclude let us precise that this work can be easily extended to mixture-models with more than two components and can be done in a nonparametric setting when using the testing procedure proposed by Butucea and Tribouley [2] as the Oracle test and the one given by Autin and Pouet [1] as the Mixing test.

An interesting extension that should really be considered is the case of mixing-weights with errors. This arises when mixing-weights are computed from a model with estimated parameters or from experts' evaluation. In this case the solution of (3) is no longer exact as the matrices Ω_X and Ω_Y are random. Preliminary simulation results tend to prove that moderate errors have a small effect.

6 Appendix

In this section we provide the technical lemmas and the proposition required to prove the asymptotic normality under interest. For the sake of simplicity, we present them with respect to X_1, \dots, X_n whereas an analogous version of them does exist for Y_1, \dots, Y_n . We recall that we assume that, for any $l \in \{1, 2\}$ the mixing-weights of the model satisfy (5).

Denote, for any $n \in \mathbb{N}^*$, any $l \in \{1, 2\}$ and any $1 \leq i \leq n$

$$W_{ni}^{(l)} = \frac{a_l(i)}{\sqrt{B_n^{(l)}}} (X_i - \omega_1(i)m_1 - \omega_2(i)m_2). \quad (15)$$

Lemma 1 *For any $1 \leq i \leq n$*

$$\mathbb{V}ar(X_i) = \sum_{l=1}^2 \left(\int \omega_l(i)(x - m_l)^2 p_l(x) dx \right) + \omega_1(i)\omega_2(i)(m_1 + m_2)^2.$$

Proof:

For any $1 \leq i \leq n$,

$$\begin{aligned}\mathbb{V}ar(X_i) &= \int (x - \omega_1(i)m_1 + \omega_2(i)m_2)^2 (\omega_1 p_1(x) + \omega_2 p_2(x)) dx \\ &= \sum_{l=1}^2 \left(\int \omega_l(i)(x - m_l)^2 p_l(x) dx \right) + \omega_1(i)\omega_2(i)(m_1 + m_2)^2.\end{aligned}$$

From Lemma 1, we immediately derive:

Lemma 2 For any $l \in \{1, 2\}$, let $B_n^{(l)}$ be defined as in (7).

$$B_n^{(l)} \geq \min(\sigma_1^2, \sigma_2^2) \sum_{i=1}^n a_l^2(i).$$

Proof:

For any $l \in \{1, 2\}$, by using Lemma 1 and the fact that, for any $1 \leq i \leq n$ $\omega_1(i) + \omega_2(i) = 1$,

$$\begin{aligned}B_n^{(l)} &= \sum_{i=1}^n a_l^2(i) \mathbb{V}ar(X_i) \\ &= \sum_{i=1}^n \left[a_l^2(i) \left(\sum_{l=1}^2 \int \omega_l(i)(x - m_l)^2 p_l(x) dx \right) + \omega_1(i)\omega_2(i)(m_1 + m_2)^2 \right] \\ &\geq \sum_{i=1}^n \left[a_l^2(i) \left(\sum_{l=1}^2 \omega_l(i) \right) \right] \min \left(\int (x - m_1)^2 p_1(x) dx, \int (x - m_2)^2 p_2(x) dx \right) \\ &= \min(\sigma_1^2, \sigma_2^2) \sum_{i=1}^n a_l^2(i).\end{aligned}$$

Lemma 3 For any $l \in \{1, 2\}$ Let $B_n^{(l)}$ and $W_{ni}^{(l)}$ ($1 \leq i \leq n$) be defined as in (7) and (15). Then, for any $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{E} \left[\left(W_{ni}^{(l)} \right)^2 \mathbf{1} \left\{ |W_{ni}^{(l)}| \geq \epsilon \right\} \right] = 0.$$

Proof:

Fix $l \in \{1, 2\}$. Let us define for any $n \in \mathbb{N}^*$

$$\begin{aligned}\kappa_n &= \min\{m_1, m_2\} + \frac{\epsilon \sqrt{B_n^{(l)}}}{\sup_i \{|a_l(i)|\}}, \\ \kappa'_n &= \max\{m_1, m_2\} - \frac{\epsilon \sqrt{B_n^{(l)}}}{\sup_i \{|a_l(i)|\}}.\end{aligned}$$

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{E} \left[\left(W_{ni}^{(l)} \right)^2 \mathbf{1} \left\{ |W_{ni}^{(l)}| \geq \epsilon \right\} \right] \\
&= \sum_{i=1}^n \int_{|y| \geq \epsilon} y^2 dF_{W_{ni}^{(l)}}(x) \\
&\leq \sum_{i=1}^n \frac{a_l^2(i)}{B_n^{(l)}} \int_{x > \kappa_n, x < \kappa'_n} (x - \omega_1(i)m_1 - \omega_2(i)m_2)^2 dF_{X_i}(x) \\
&\leq 2 \sum_{i=1}^n \frac{a_l^2(i)}{B_n^{(l)}} \left[\sum_{l \in \{1,2\}} \omega_l(i)^3 \int_{x > \kappa_n, x < \kappa'_n} (x - m_l)^2 p_l(x) dx \right. \\
&\quad \left. + \omega_2^2(i)\omega_1(i) \int_{x > \kappa_n, x < \kappa'_n} (x - m_2)^2 p_1(x) dx + \omega_1^2(i)\omega_2(i) \int_{x > \kappa_n, x < \kappa'_n} (x - m_1)^2 p_2(x) dx \right] \\
&\leq 2 \left(\sum_{i=1}^n \frac{a_l^2(i)}{B_n^{(l)}} \right) \sup_{n \in \mathbb{N}^*} \left(\sum_{(k,l) \in \{1,2\}^2} \int_{x > \kappa_n, x < \kappa'_n} (x - m_k)^2 p_l(x) dx \right).
\end{aligned}$$

Using Lemma 2, for any $n \in \mathbb{N}^*$,

$$\sum_{i=1}^n \frac{a_l^2(i)}{B_n^{(l)}} \leq (\min(\sigma_1^2, \sigma_2^2))^{-1}.$$

Then, since variances under p_1 and p_2 are finite, the supremum over n tends to 0 in the integrals above, according to Lebesgue dominated convergence theorem.

Lemma 4 For any $1 \leq i \leq n$,

$$\mathbb{E} [(X_i - \mathbb{E}(X_i))^4] \leq C(m_1, m_2, p_1, p_2),$$

where $C(m_1, m_2, p_1, p_2) := 32 \max_{(k,l) \in \{1,2\}^2} \int (x - m_k)^4 p_l(x) dx$.

Proof:

We have

$$\begin{aligned}
& \mathbb{E} [(X_i - \mathbb{E}(X_i))^4] \\
&= \mathbb{E} [(X_i - \omega_1(i)m_1 - \omega_2(i)m_2)^4] \\
&\leq 8 \omega_1(i)^4 \mathbb{E} ((X_i - m_1)^4) + 8 \omega_2(i)^4 \mathbb{E} ((X_i - m_2)^4) \\
&= 8 \left[\omega_1(i)^5 \int (x - m_1)^4 p_1(x) dx + \omega_1(i)^4 \omega_2(i) \int (x - m_1)^4 p_2(x) dx \right. \\
&\quad \left. + \omega_2(i)^4 \omega_1(i) \int (x - m_2)^4 p_1(x) dx + \omega_2(i)^5 \int (x - m_2)^4 p_2(x) dx \right] \\
&\leq 32 \max_{(k,l) \in \{1,2\}^2} \int (x - m_k)^4 p_l(x) dx.
\end{aligned}$$

Lemma 5 For any $1 \leq l \leq 2$, the estimator $\hat{m}_l = \frac{1}{n} \sum_{i=1}^n a_l(i) X_i$ of m_l is consistent, that is

$$\hat{m}_l \xrightarrow{\text{Proba}} m_l.$$

Proof:

Let $\epsilon > 0$ and $l \in \{1, 2\}$. We have, using Bienayme-Chebyshev inequality and Lemma 1

$$\begin{aligned}
& \mathbb{P} (|\hat{m}_l - m_l| > \epsilon) \\
&\leq \frac{1}{n^2 \epsilon^2} \sum_{i=1}^n a_l^2(i) \mathbb{V}ar(X_i) \\
&= \frac{1}{n^2 \epsilon^2} \sum_{i=1}^n a_l^2(i) \left[\sum_{l \in \{1,2\}} \left(\omega_l(i) \sigma_l^2 + \omega_l(i)(1 - \omega_l(i)) \frac{(m_1 + m_2)^2}{2} \right) \right] \\
&\leq \frac{2}{n \epsilon^2 K} \left[\sum_{l \in \{1,2\}} \sigma_l^2 + \frac{(m_1 + m_2)^2}{4} \right].
\end{aligned}$$

Last inequality is obtained by using assumption on the smallest eigenvalue of $\Omega' \Omega$, that is larger than Kn (with $K > 0$) and the fact that the supremum value for $x \in [0, 1]$ of $x \rightarrow x(1 - x)$ is equal to $\frac{1}{4}$. The right-hand side clearly tends to 0 when n goes to infinity. We conclude that \hat{m}_l is consistent.

Proposition 1 For any $l \in \{1, 2\}$, any $n \in \mathbb{N}^*$ and any $1 \leq i \leq n$ consider $W_{ni}^{(l)}$, defined as in (15).

$$\sum_{i=1}^n W_{ni}^{(l)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Proof:

We apply Theorem 4.2 in Petrov [6]. It is the general setup for the Central Limit Theorem, for the triangular array of series $(W_{ni})_{i,n}$ of independent random variables X_i (that are not identically distributed).

If the three conditions are satisfied for any $\epsilon > 0$ and any $\tau > 0$

1. $\lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(|W_{ni}^{(l)}| \geq \epsilon) = 0,$
2. $\lim_{n \rightarrow \infty} \sum_{i=1}^n \int_{|y| < \tau} y dF_{W_{ni}^{(l)}}(y) = 0,$
3. $\lim_{n \rightarrow \infty} \sum_{i=1}^n \left\{ \int_{|y| < \tau} y^2 dF_{W_{ni}^{(l)}}(y) - \left(\int_{|y| < \tau} y dF_{W_{ni}^{(l)}}(y) \right)^2 \right\} = 1,$

then $\sum_{i=1}^n W_{ni}^{(l)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$

Let us prove that the three conditions are satisfied. Let $\epsilon > 0$. Using Bienayme-Chebyshev inequality,

$$\sum_{i=1}^n \mathbb{P}(|W_{ni}^{(l)}| \geq \epsilon) \leq \epsilon^{-2} \sum_{i=1}^n \mathbb{E} \left[\left(W_{ni}^{(l)} \right)^2 \mathbf{1}_{\{|W_{ni}^{(l)}| \geq \epsilon\}} \right].$$

Hence, condition 1 is clearly satisfied by using Lemma 3.

Let us move to condition 2. We use the same trick as above. For any $\tau > 0$

$$\sum_{i=1}^n \int_{|y| < \tau} y dF_{W_{ni}^{(l)}}(y) = \sum_{i=1}^n \left[\int y dF_{W_{ni}^{(l)}}(y) - \int_{|y| \geq \tau} y dF_{W_{ni}^{(l)}}(y) \right].$$

The first summand is equal to 0 as the variables $W_{ni}^{(l)}$ are centered.

$$\begin{aligned} \sum_{i=1}^n \left| \int_{|y| \geq \tau} y dF_{W_{ni}^{(l)}}(y) \right| &\leq \sum_{i=1}^n \int_{|y| \geq \tau} |y| dF_{W_{ni}^{(l)}}(y) \\ &\leq \tau^{-1} \sum_{i=1}^n \mathbb{E} \left[\left(W_{ni}^{(l)} \right)^2 \mathbf{1}_{\{|W_{ni}^{(l)}| \geq \tau\}} \right]. \end{aligned}$$

Condition 2 is clearly satisfied by using Lemma 3.

We end the proof with condition 3. There are two parts (because of two summands) in this condition. For the first part we proceed exactly as in condition

2. Indeed we have

$$\sum_{i=1}^n \int_{|y|<\tau} y^2 dF_{W_{ni}^{(l)}}(y) = \sum_{i=1}^n \int y^2 dF_{W_{ni}^{(l)}}(y) - \sum_{i=1}^n \int_{|y|\geq\tau} y^2 dF_{W_{ni}^{(l)}}(y).$$

The first summand is exactly equal to 1 and the second one tends to 0 as n goes to infinity, according to Lemma 3. Therefore it remains to prove that the second part tends to 0 when n goes to infinity. Because the variables $W_{ni}^{(l)}$ are centered and according to Cauchy-Schwarz inequality:

$$\begin{aligned} \sum_{i=1}^n \left(\int_{|y|<\tau} y dF_{W_{ni}^{(l)}}(y) \right)^2 &= \sum_{i=1}^n \left(\int_{|y|\geq\tau} y dF_{W_{ni}^{(l)}}(y) \right)^2 \\ &\leq \sum_{i=1}^n \int_{|y|\geq\tau} y^2 dF_{W_{ni}^{(l)}}(y) \\ &= \sum_{i=1}^n \mathbb{E} \left[\left(W_{ni}^{(l)} \right)^2 \mathbf{1}_{\{|W_{ni}^{(l)}| \geq \tau\}} \right]. \end{aligned}$$

Still using Lemma 3, we conclude that the second part we are interested in tends to 0 when n goes to infinity, as expected.

References

- [1] Autin, F., and Pouet., C. (2011). Test on components of densities mixture. To appear in *Statistics & Risk Modeling*.
- [2] Butucea, C., and Tribouley, K. (2006). Nonparametric homogeneity tests. *Journal of Statist. Plann. Inference*, **136**, 597-639.
- [3] Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.*, **60**, 549-576.
- [4] McLachlan G.J., and Peel, D. (2000). *Finite mixture-models*. Wiley, New York.
- [5] Maiboroda, R.E. (2000). An asymptotically effective estimate for a distribution from a sample with a varying mixture. *Theory Probab. Math. Statist.*, **61**, 121-130.
- [6] Petrov, V.V. (1995). *Limit theorems of probability theory*. Oxford University Press.
- [7] Pokhyl'ko, D. (2005). Wavelet estimators of a density constructed from observations of a mixture. *Theor. Prob. and Math. Statist.*, **70**, 135-145.
- [8] U.S. Census Bureau (2006). American Community Survey. Profile: California, Illinois and New York (Available from <http://www.census.gov/>).

- [9] Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**, 2835.